


A Comparative Study of Various Transfer Learning Models on Skin Cancer Confirmation Methods


Mehmet Ali Altuncu

(Department of Computer Engineering, Faculty of Engineering, Kocaeli University, Kocaeli, Turkey)

 <https://orcid.org/0000-0002-2948-3937>, mehmetali.altuncu@kocaeli.edu.tr)


Kaplan Kaplan

(Department of Software Engineering, Faculty of Engineering, Kocaeli University, Kocaeli, Turkey)

 <https://orcid.org/0000-0001-8036-1145>, kaplan.kaplan@kocaeli.edu.tr)

Melih Kuncan*

(Department of Electrical and Electronics Engineering, Faculty of Engineering, Siirt University, Siirt, Turkey)

 <https://orcid.org/0000-0002-9749-0418>, melihkuncan@siirt.edu.tr)

Abstract: Skin cancer confirmation is critical in determining a patient's treatment planning process after diagnosis. A proper confirmation process enables the determination of the type, stage, and other characteristics of skin cancer, helping to plan the appropriate treatment. These methods prevent the progression of the disease, thereby contributing to a better response to treatment and improving the patient's quality of life. Dermoscopic images are commonly used to confirm skin cancer types. To obtain meaningful results from these images, researchers often apply artificial intelligence techniques in such studies. Specifically, transfer learning models have been commonly used to enhance the features of these images due to the limited availability of medical image data and the difficulty in extracting meaningful information from such data. While most studies focus on classifying skin cancer types, this research aims to classify skin cancer confirmation types using dermoscopic dataset images. Dermoscopic HAM10000 dataset images were used for this purpose. The dataset includes four different confirmation methods: confocal, consensus, follow-up, and histopathology. Four distinct transfer learning models—Resnet-50, Resnet-101, VGG19, and InceptionResnetV2—were utilized. Additionally, ensemble learning was conducted based on the results of these models using the maximum voting approach. The highest success rate was achieved with Resnet-101 at 96.04%. Considering the comparative results, the accuracy of our promising model proved to be significantly high.

Keywords: Classification, skin cancer, transfer learning, diagnosis, deep learning

Categories: I.4.7, I.5, I.6.4, I.2.6

DOI: 10.3897/jucs.118220

1 Introduction

Skin cancer is one of the most common diseases in humans with white skin. It is defined as a category that is bad for skin cancer. In addition, the incidence of skin cancer is increasing rapidly around the world. Melanoma is defined as one of the deadliest types of skin cancer, and it has been observed from literature studies that approximately

50,000 people die from melanoma worldwide annually. This death rate is approximately 0.7% of all cancer deaths. However, when the death rates are analyzed by country, they show significant differences. In the 10-year period from 2008 to 2018, the annual number of melanoma cases increased by more than 50% because of the increase in ultraviolet rays. Although melanoma is known as one of the deadliest types of skin cancer, early diagnosis significantly affects the survival rate [Bozkurt, 2023]. There are several types of confirmation for a clear diagnosis of skin cancer. These can be broadly divided into 4 categories. The first step is the examination of the lesions under a microscope using various methods. This is called histopathology (histo). The second method is to diagnose the disease by real or follow-up examination. The third method is to make a diagnosis by an expert consensus. The last method in disease diagnosis is the use of *in vivo* confocal microscopy (confocal), which is common in fluorescent imaging [Tschandl et al., 2018]. The first step in diagnosing a malignant lesion by a dermatologist is visual examination of the suspicious skin area. Accurate diagnosis is crucial because of the similarities between different types of lesions. In addition, diagnostic accuracy is directly related to the professional experience of the specialist medical doctor [Brinker et al., 2018]. Without additional technical support from literature searches, dermatologists generally have an accuracy rate of 65%–80% in diagnosing melanoma [Nami et al, 2012, Fabbrocini et al., 2010, Haenssle et al., 2018, Demir 2021, Argenziano and Soyer, 2001].

In unclear situations, visual-based examination is supported by dermoscopic images taken with a very high sensitivity and high-resolution magnifying camera, which is a result of developing technological innovations. In these cases, since the lighting has a very important place during the recording of the pictures (images) to be taken, light, lighting, etc. processes should be controlled and their reflections on the skin, etc. A high-specification filter is used to minimize adverse events. In this way, it is possible to clearly see the deeper skin layers in the body. With the abovementioned light, lighting, filter, etc., it is stated that the accuracy of skin lesion diagnosis can be increased by approximately 50% by adjusting the conditions in the most optimal way. Thanks to the integration of visual analysis and dermoscopic images, it has been observed by expert dermatologists that an absolute melanoma detection accuracy of approximately 75%–85% is achieved.

Different levels of expertise among specialists can lead to significant differences in diagnostic accuracy. As a result, there is considerable interest in screening initiatives and the development of partially or completely automated computer-aided diagnostic systems that serve as a second, independent opinion [Bozkurt, 2023]. Examining the most popular methods, particularly in computer-based diagnostic systems, reveals that artificial intelligence models rank highest. [Kittler et al, 2002, Ali and Deserno, 2012, Fabbrocini et al, 2011, Krizhevsky et al., 2017].

In addition, in recent years, there have been different studies in the field of computer-based medicine in different fields. Researchers have recently begun to pay particular attention to biomedical data. In the field of medicine, the use of electronic devices (sensors, measuring devices, new technological equipment, etc.) has become widespread in diagnosis and treatment studies. Studies in the field of medicine are widely concentrated on the diagnosis of diseases. These diseases are Covid-19, heart diseases, muscle diseases, different types of cancer, etc. disease diagnoses are widely preferred by researchers. In recent years, researchers have been continuing their studies on skin cancer at a significantly faster pace than these disease diagnosis studies.

2 Literature Review

In recent years, different computer-based signal processing studies have been conducted in different fields. The main purpose of signal processing is to obtain information from signals or to make predictions using computer approaches. Researchers have recently begun to pay particular attention to biomedical signals. Many different applications are widely used in different fields (medicine, sports, security, etc.) by using medical signals in both academic and research project studies. Especially recently, rapid technological innovations have greatly affected medical science as well as all science and application fields. As a result, the use of electronic devices (sensors, measuring devices, new technological equipment, etc.) in diagnosis and treatment studies in the field of medicine has become widespread. Studies in the field of medicine are widely concentrated on the diagnosis of diseases. These diseases include Covid-19, heart diseases, muscle diseases, different types of cancer, etc. An example can be given. Disease diagnoses are widely preferred by researchers. In recent years, researchers have been continuing their studies on skin cancer at a significantly faster pace than these disease diagnosis studies.

[Rahi et al., 2021] proposed a semi-automatic classification system for skin cancer detection using transfer learning models. In this study, transfer learning architectures such as Mobilenet V1, Inception-V3, VGG16, VGG19, and U-Net were compared. During the tests, the highest accuracy of 91% of U-Net, which is a convolutional network, was obtained in the HAM10000 dataset.

[Huang et al., 2021] proposed a deep learning model that can run on mobile devices and cloud platforms for remote diagnosis applications in skin cancer classification. In the study, training data were pre-processed for normalization, dithering, vertical flip, and horizontal flip. The DenseNet algorithm, which is a CNN-based method, was used as a classifier for skin cancer detection. Because of the study, 85.8% accuracy was obtained in the HAM10000 dataset.

[Khan et al., 2021a] proposed a two-stage deep learning architecture for segmentation and classification of skin lesions. In the first step, a mask RCNN model is proposed for segmentation, and in the second step, a CNN architecture with six convolutional and one fully connected layer is proposed. An accuracy of 86.5% was obtained in the HAM10000 dataset used for classification.

[Popescu et al., 2022] proposed a system for the detection of skin lesions, including multi-convolutional neural networks trained on the HAM10000 dataset. In the proposed system, the decision-making process is carried out with ensemble learning logic by combining the advantages of different CNN networks. Data augmentation techniques were also used to balance the dataset. The proposed system achieved an average of 86.71% accuracy in detecting 7 different skin lesions.

[Srinivasu et al., 2021] again used the HAM10000 dataset to classify skin diseases. In the study, the separation of the dataset as training and testing was performed randomly. In this study, data augmentation techniques were used to balance the dataset. Based on MobileNet V2 and the LSTM approach, the proposed model was compared with traditional methods such as CNN, FTNN, and HARIS, and it was stated that it gave better results with 85.34% accuracy.

[Hameed et al., 2021] proposed data augmentation and multiple image processing techniques. In this study, techniques such as filtering and histogram equalization were used to detect the contaminated area in the images. The proposed model is based on a

stacked CNN, similar to classical deep learning models. Therefore, image enlargement methods were used to increase the training data set, and 95.2% accuracy was obtained.

[Thurnhofer-Hemsi and Domínguez, 2021] applied the transfer learning method to five convolutional neural networks (DenseNet201, GoogLeNet, Inception-ResNetV2, InceptionV3, MobileNetV2). The proposed method consists of two stages to deal with class imbalance. In the first stage, differentiation of the images with and without the non-nevi was performed, and in the second stage, classification of the non-nevi types was performed. The experiments indicated that DenseNet201 was the best deep network in the HAM10000 dataset with 95.09% accuracy.

[Alam et al., 2022] compared deep learning-based models AlexNet, Inception-V3, and RegNetY-320 algorithms to classify skin cancer. In this study, the HAM10000 dataset was used both in its original (unbalanced) form and in its balanced form using data augmentation methods such as rotating and flipping. The results indicated that RegNetY-320 gave the best results in both cases. While the accuracy was 85% on the unbalanced dataset, the performance increased to 91% after applying the proposed framework.

[Khan et al., 2021b] proposed a deep learning-based method for multiclass skin lesion segmentation and classification. In the proposed method, the input images are developed using local color-controlled histogram intensity values. For feature extraction, we provide a pre-trained deep CNN model. In this study, ISBI 2016, ISBI 2017, ISBI 2018, and PH2 datasets were used for segmentation, and HAM10000 datasets were used for multiclass skin lesion classification. The efficiency of the method was compared with the performances of different neural networks, and 90.67% accuracy was obtained in the classification process.

[Hoang et al., 2022] proposed a two-stage method for skin lesion classification. We used entropy-based weighting and first-order cumulative moment of the skin image to distinguish the lesion from the background in the first step. In the second step, they applied a two-dimensional ShuffleNet network to classify the image. In the classification phase, 86.33% accuracy was obtained in the HAM10000 dataset.

[Garg et al., 2021] proposed an architecture that uses image processing and deep learning together to classify images in the HAM10000 dataset. The proposed architecture is based on feature extraction and training using the transfer learning model of these features. In this study, the number of images was increased using various image enlargement techniques. The results were compared with the XGBoost, SVM, and Random Forest algorithms, and it was stated that the proposed transfer learning approach gave better results with an accuracy of 90.51%.

[Yu and Marganec, 2021] proposed an IoT-based architecture for remote skin disease diagnostic applications. The proposed architectural color feature comprises model transfer and data balancing stages. The HAM10000 dataset was used to measure the performance of the system, and VGG16, Inception, Xception, MobileNet, ResNet-50, and DenseNet161 deep learning models were compared. It was stated that DenseNet161 was more successful than the other models with 86.5% accuracy.

[Khan et al., 2021c] proposed a new approach aimed at diagnosing the disease by capturing images of skin lesions using a mobile device. First, image segmentation was performed using CNN architecture and high-dimensional contrast transformation. In the second stage, classification was performed with the help of DenseNet CNN and multi-class ELM classifier. With the proposed model, an accuracy of 88.39% was obtained in the HAM10000 dataset.

[Lan et al., 2022] proposed a capsular network called FixCaps, which reduces the number of nuclei in the convolutional layer for classification of skin lesions. The authors stated that owing to their proposed capsule network, they achieved less computational cost and a higher success rate (96.49% accuracy on the HAM10000 dataset) compared to standard CNN models.

[Datta et al., 2021] proposed the Soft-Attention mechanism to increase the value of important features in dermoscopic images and eliminate noise-causing features. In this study, the performance of VGG, ResNet, InceptionResnetV2, and DenseNet architectures were compared with models with and without the Soft-Attention mechanism to classify skin lesions. The authors stated that models with the Soft-Attention mechanism performed an average of 3% better and achieved 93.4% accuracy.

[Ren, 2023] compared transfer learning methods for the detection of monkeypox. In the study, twelve transfer learning methods were applied, and it was stated that the best results were obtained with the DenseNet201 model in both binary and multi-class classification.

[Bibi et al., 2023] proposed a system for multi-class skin cancer detection consisting of image preprocessing, feature extraction and selection, and classification steps. In the proposed system, residual blocks were added to the final layer of the DarkNet-53 and DenseNet201 models in the classification stage, which was stated to improve the success rate.

[Maqsood and Damaševičius, 2023] have proposed a hybrid method based on deep learning. In the proposed method, a CNN architecture is used for segmentation, transfer learning methods are employed during the training phase, and a multi-class support vector machine (MC-SVM) is utilized for classification, resulting in high performance.

[Hussain et al., 2023] have proposed a system consisting of image preprocessing and classification stages for multi-class skin lesion classification. In the study, hyperparameter selection in transfer learning methods was automatically determined using genetic algorithm. It has been noted that this led to an increase in the learning rate. When the literature is examined, it has been seen that studies on skin cancer are generally conducted on the classification of skin cancer types. The confirmation of these cancer types is of great importance for definitive cancer diagnosis. Within the scope of this study, classification of confirmation types of skin cancer disease with transfer learning approaches was performed using HAM10000 skin cancer data. It is thought that the proposed approach can also be used in the classification of different medical signals.

3 Material and Method

In this study, transfer learning models were used to classify skin cancer confirmation types. These models are Resnet-50, Resnet-101, VGG19, and InceptionResnetV2, which are used effectively in the literature.

3.1 Resnet-50

ResNet-50 is a deep convolutional neural network (CNN) trained on the ImageNet dataset. It is a variant of the ResNet architecture introduced by [He et al., 2016a] in the article "Deep Residual Learning for Image Recognition". ResNet-50 is known for its high performance and generalization ability in a wide variety of image classification

tasks. It has been widely used as a basic model in various computer vision applications and has shown outstanding results for a variety of criteria. Using residual connections, a core component of the ResNet architecture, has enabled the network to learn more complex connections as well as identity functions. This has helped improve very deep network training, which is normally hampered by disappearing gradients and below-average performance. An ImageNet dataset containing more than one million real images and 1000 classifications assigned to them was used to train ResNet-50 [He et al., 2016a]. The model has already been trained on this dataset and is adjusted for use in other image classification applications by adding a few more layers and retraining on the new dataset. The architecture of the Resnet-50 model is shown in Figure 1.

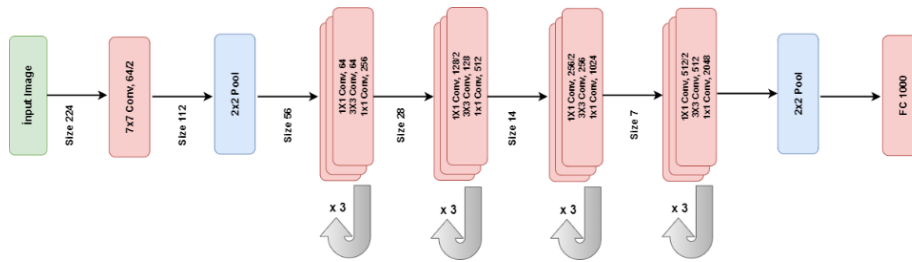


Figure 1: Resnet-50 model architecture

3.2 Resnet-101

The ResNet-101 architecture is also presented in the publication "Deep Residual Learning for Image Recognition" by [He et al., 2016]. It is widely used in many computer vision applications and has produced superior results for a variety of criteria [Chen et al., 2017]. ResNet-101 has a deep network design with 101 layers organized into a series of residual blocks. Each residual block consists of multiple convolutional layers and shortcut paths used to learn both the identity and more complex functions. The total network is created by stacking these residual blocks on top of each other, with a global average pooling layer and a fully connected layer that produces the final output. Using residual links allows the network to learn more easily and train deeper networks without suffering from degradation, where deep networks tend to perform worse than shallower networks as depth increases. In addition to residual connections, the ResNet architecture uses batch normalization and ReLU enable functions to aid deep network training. Overall, the ResNet architecture is intended to be more efficient and easier to train than classical CNNs and has shown superior performance in a range of image classification tasks [He et al., 2016b].

3.3 VGG-19

VGG19 is a convolutional neural network model developed by the Visual Geometry Group (VGG) at the University of Oxford. The model was developed by Karen Simonyan and Andrew Zisserman, researchers at VGG [Simonyan and Zisserman, 2014]. It is a 19-layer deep learning model, including convolutional, max pooling, and fully connected layers. VGG19 is trained on the ImageNet dataset consisting of more

than 1 million images belonging to 1000 different classes [Stateczny et al., 2022]. The model can use the images as "dog", "cat", "car". It is designed to classify into one of these classes. Due to its efficiency and simplicity, the VGG19 model is now frequently used in image classification applications. It has served as a base model in various computer vision tasks and has been cited in a great deal of scientific literature. VGG19 and its derivatives have been used in tasks such as object identification and segmentation, as well as image classification [Litjens et al., 2017]. Thirteen convolutional layers and three fully connected layers make up the 19 layers of the VGG19 model. Convolutional layers use a number of filters to identify specific patterns or characteristics in the input image. Maximum pooling layers reduce the size of the feature maps generated by convolutional layers while preserving important information. The fully connected layers classify the input image using the outputs from the convolutional and max pooling layers. A large dataset of annotated images was used to train the model, enabling it to learn the relationships between the input features and their corresponding labels. During training, the model's parameters are adjusted to minimize the error between the predicted and actual labels. Once trained, the model can identify new images by utilizing the learned filters to extract features and then predicting the image's label with the fully connected layers [Jaworek-Korjakowska et al., 2019]. Figure 2 depicts the architecture of the VGG19 model.

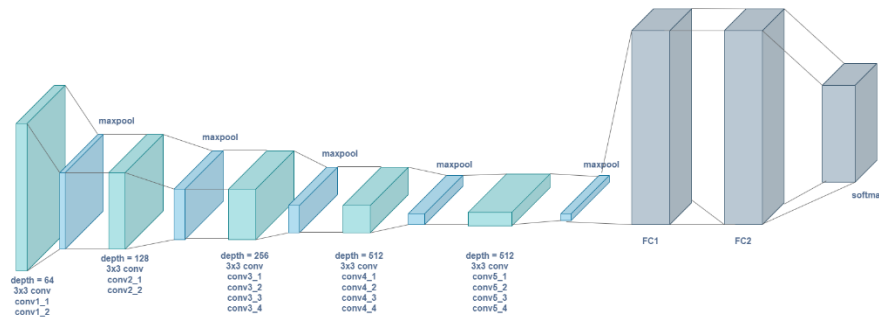


Figure 2: VGG19 model architecture

3.4 InceptionResnetV2

InceptionResNetV2 is a convolutional neural network model developed by Google for the ImageNet Large-Scale Visual Recognition Competition (ILSVRC) [Szegedy et al., 2017]. This is a variation of the Inception model developed by Google for the same competition and combines the Inception architecture with the residual connections used in the ResNet model. It works by extracting features from the input images using a series of convolutional and maximum pooling layers and then using these features to classify the images into the appropriate category. The InceptionResNetV2 model consists of more than 100 layers, including convolutional, maximum pooling, and fully connected layers. Convolutional layers apply a series of filters to the input image to detect certain patterns or features in the image. Maximum pooling layers reduce the size of feature maps produced by convolutional layers while protecting important information. InceptionResNetV2 is part of the Inception model family, which is known for its ability to achieve good performance in image classification tasks while being

relatively efficient in terms of the number of parameters required and computational resources [Szegedy et al., 2017]. The InceptionResNetV2 model has been extensively studied and referenced in the academic literature and has been used as the baseline model in many computer vision tasks [Wang et al., 2021]. A schematic of the InceptionResNetV2 model is shown in Figure 3.

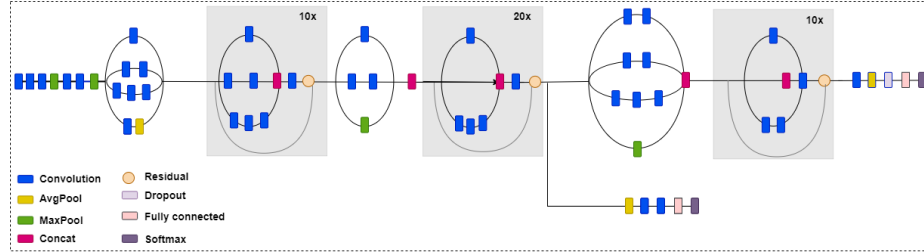


Figure 3: Schematic diagram of the InceptionResNetV2 model

3.5 Performance Metrics

The performance of the methods proposed in the study was evaluated using 5 different metrics: accuracy, precision, recall, and F-score. Four parameters called true positive (TP), true negative (TN), false positive (FP), and false negative (FN) are used to define these metrics (Equation 1-4). TP refers to the number of samples correctly classified in the skin cancer dataset, TN denotes the number of correctly classified normal samples, FP is the number of wrongly classified normal samples, and FN shows the number of wrongly classified samples.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (1)$$

$$\text{Precision} = \text{TN} / (\text{TN} + \text{FP}) \quad (2)$$

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN}) \quad (3)$$

$$f1 - \text{Score} = (2 * \text{Recall} * \text{Precision}) / (\text{Recall} + \text{Precision}) \quad (4)$$

The Receiver Operator Characteristics) curve is employed for visualizing the performance of models in classification problems. The FP ratio of the ROC curve is on the x-axis and the TP ratio is on the y-axis. The FP rate can be found from Equation 5 and the TP rate from Equation 6. When the ROC curve shifts to the upper left corner, better classification is performed. When the curve approaches the center, classification performance deteriorates [Park et al., 2004].

$$\text{FP Rate} = \frac{\text{FP}}{\text{TN} + \text{FP}} \quad (5)$$

$$\text{TP Rate} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (6)$$

4 Dataset

In this study, the HAM10000 dataset [Tschandl et al., 2018] was used to analyze the experimental results and proposed approaches. In both the training and testing phases, the records in the data were used in their original form without any changes. The HAM10000 dataset is one of the most common datasets used in skin lesion classification. The dataset contains 10015 dermoscopic images collected from the dermatology department of the Medical University of Viena (Austria). The HAM10000 dataset has four classes for skin cancer confirmation: histo, follow-up, consensus, and confocal. The distribution of these classes is shown in Table 1.

Class	Percentage (%)
histo	53
follow up	37
consensus	9
confocal	1

Table 1: Distribution of classes in the HAM1000 dataset

5 Experimental Results

In this study, four different transfer learning models were used: Resnet-50, Resnet-101, VGG19, and models. 70% of the entire dataset was used in the training phase and 30% in the testing phase. In other words, 7010 of 10015 dermoscopic image data were used for educational purposes and 3005 for testing purposes. The training and test data were chosen randomly from the data. The same data were used in the training and testing stages of each model to measure the comparative success of the models and for a fair classification experiment. Confusion matrices were created at each model testing stage, and performance measures were calculated. As data tags, '1' represents 'Confocal' class, '2' represents 'Consensus' class, '3' represents 'follow_up' class, and '4' represents 'histo' class.

These images, which originally had 600×450 pixels, were converted to 224×224 in the pre-processing stage. Then, the model hyperparameters were selected and the training and testing stages of the models were performed.

5.1 Results obtained using the VGG19 model

The training stage was performed using the hyperparameters in Table 2 of the VGG19 model used in the classification. Figure 4 shows the success and loss functions obtained during the training. The success and loss functions show that there is no excessive memorization for this model and that a successful training phase is performed.

Hyperparameter	Value
Optimizer	Adam
Initial Learning Rate	0.00001
Batch Size	64
Max Epochs	10
L2 Regularization	0.0005
Optimizer	Adam

Table 2: Hyperparameters used for VGG19 model

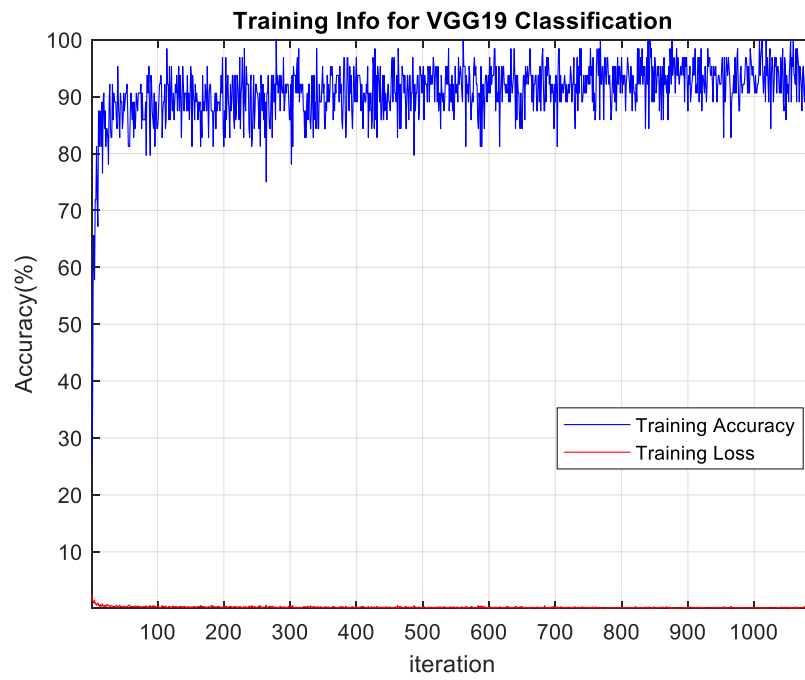


Figure 4: Accuracy and loss function obtained during the VGG19 training phase

In Figure 5, the confusion matrix obtained from the test data after training the VGG19 model is given.

True Class	1	12	1		8
	2	8	79	17	167
	3		3	1107	1
	4	6	19	19	1558
		1	2	3	4
		Predicted Class			

Figure 5: Confusion matrix for VGG19

The performance criteria obtained because of VGG19 are given in Table 3. The model achieved an accuracy rate of 91.71% in total performance. When the performance values were examined, the VGG19 model achieved high success rates for the 3rd and 4th grades. It was observed that the expected performance could not be achieved because of the low class size of the success rates of the other classes.

Class	Recall	Precision	F-Score
1	0.4615	0.5714	0.5106
2	0.7745	0.2915	0.4235
3	0.9685	0.9964	0.9822
4	0.8985	0.9725	0.9340
Weighted Average	0.9101	0.9171	0.9028

Table 3: Evaluation metrics for VGG19

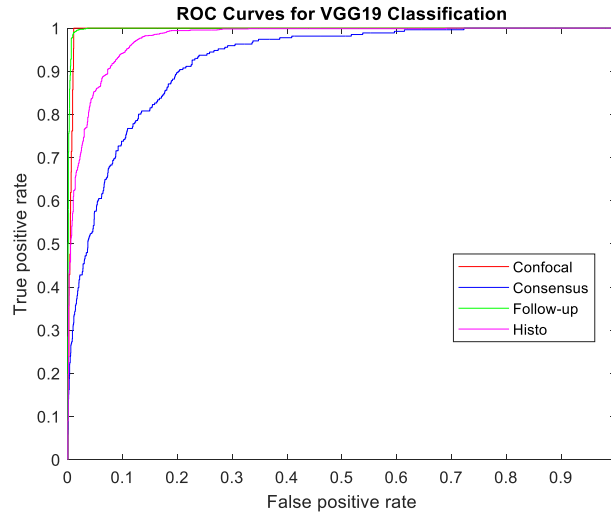


Figure 6: ROC curve for VGG19

The ROC curve for the VGG 19 model is shown in Figure 6. When the ROC curve is analyzed, a successful classification is observed because all classification curves are close to the upper margin. The most unsuccessful classification among the classes was obtained in the “Consensus” class.

5.2 Results obtained using the Resnet-50 model

The training stage was performed using the parameters in Table 4 of the Resnet-50 model used in the classification. Figure 7 shows the success and loss functions obtained during the training. The success and loss functions show that there is no excessive memorization for this model and that a successful training phase is performed.

Hyperparameter	Value
Optimizer	rmsprop
Initial Learning Rate	0.00001
Batch Size	64
Max Epochs	10
L2 Regularization	0.0005

Table 4: Hyperparameters used for Resnet-50

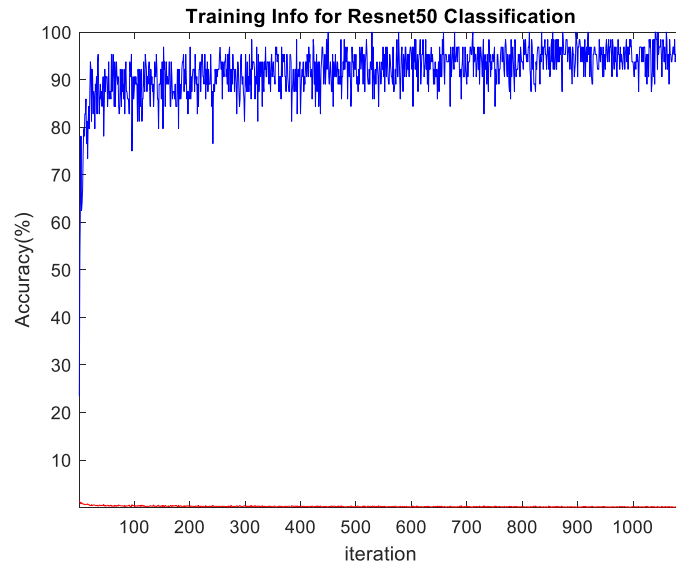


Figure 7: Accuracy and loss function obtained during the Resnet-50 training phase

In Figure 8, the confusion matrix obtained from the test data after the training phase with the Resnet-50 model is given.

1	13	5		3
2	9	131	10	121
3		6	1097	8
4	6	46	11	1539
	1	2	3	4

Predicted Class

Figure 8: Confusion matrix for Resnet-50

The performance criteria obtained from Resnet-50 are given in Table 5. The model achieved an accuracy rate of 92.51% in total performance. When Table 5 is examined, the Resnet-50 model achieved high success rates for the 3rd and 4th grades. The expected performance could not be achieved because the success rates of the other classes were small.

Class	Recall	Precision	F-Score
1	0.4642	0.6190	0.5305
2	0.6968	0.4833	0.5707
3	0.9812	0.9874	0.9842
4	0.9210	0.9606	0.9403
Weighted Avg.	0.9198	0.9250	0.9203

Table 5: Evaluation metrics for Resnet-50

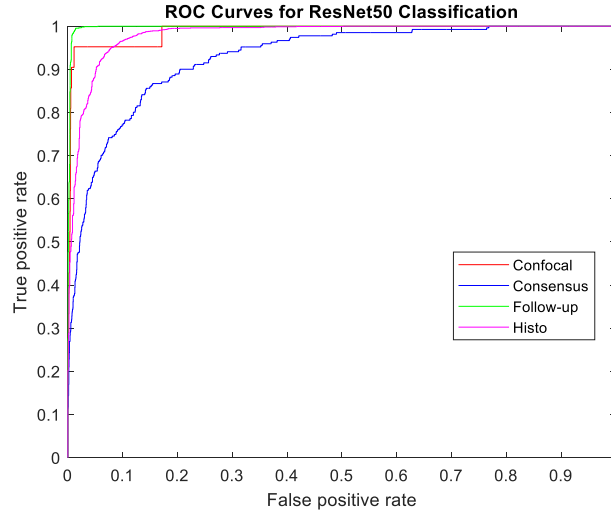


Figure 9: ROC curve for the Resnet-50 classification

Figure 9 shows the ROC curve for the Resnet-50 model. When the ROC curve is analyzed, a successful classification is observed because all classification curves are close to the upper margin. The most unsuccessful classification among the classes was obtained in the “Consensus” class.

5.3 Results obtained using the InceptionResnetV2 model

The training stage was performed using the parameters in Table 6 of the InceptionResnetV2 model used in the classification. Figure 10 shows the success and loss functions obtained during the training. The success and loss functions show that there is no excessive memorization for this model and that a successful training phase is performed.

Hyperparameter	Value
Optimizer	rmsprop
Initial Learning Rate	0.00001
Batch Size	32
Max Epochs	10
L2 Regularization	0.05

Table 6: Hyperparameters used in InceptionResnetV2

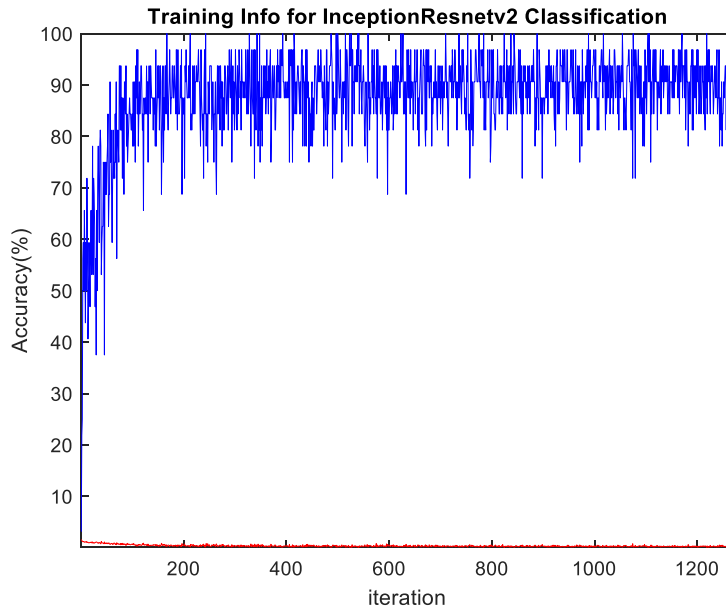


Figure 10: Accuracy and loss function obtained in the InceptionResnetV2 training phase

In Figure 11, the confusion matrix obtained from the test data after training the InceptionResnetV2 model is given.

		1		20	
1		1			
	2	68	17	186	
2					
	3	7	1097	7	
3					
	4	47	12	1543	
4					
		1	2	3	4
	True Class	Predicted Class			

Figure 11: Confusion matrix for Inception Resnet V2

The performance criteria obtained from InceptionResnetV2 are given in Table 7. The model achieved an accuracy rate of 90.12% in total performance. When Table 9 is examined, the values belonging to the first class could not be calculated because of not

being able to classify the data correctly with this model. Looking at the other classes, high achievements were obtained for the 3rd and 4th grades, where the amount of data belonging to the class was large.

Class	Recall	Precision	F-Score
1	-	0	-
2	0.5528	0.2509	0.3451
3	0.9742	0.9874	0.9807
4	0.8787	0.9631	0.9189
Weighted Avg.	0.8784	0.9011	0.8835

Table 7: Evaluation metrics for Inception Resnet V2

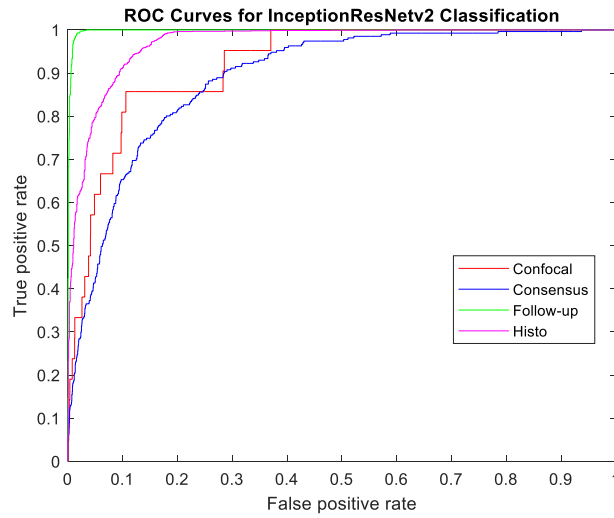


Figure 12: ROC curve for the InceptionResnetV2 classification

Figure 12 shows the ROC curve for the InceptionResnetV2 model. When the ROC curve is analyzed, a successful classification is observed because all classification curves are close to the upper margin. The most unsuccessful classification among the classes was obtained in the “Consensus” class.

5.4 Results obtained using the Resnet-101 model

The training stage was performed using the parameters in Table 8 of the Resnet-101 model used in the classification. Figure 13 shows the success and loss functions obtained during the training. The success and loss functions show that there is no excessive memorization for this model and that a successful training phase is performed.

Hyperparameter	Value
Optimizer	rmsprop
Initial Learning Rate	0.00001
Batch Size	64
Max Epochs	10
L2 Regularization	0.0005

Table 8: Hyperparameters used for Resnet-101

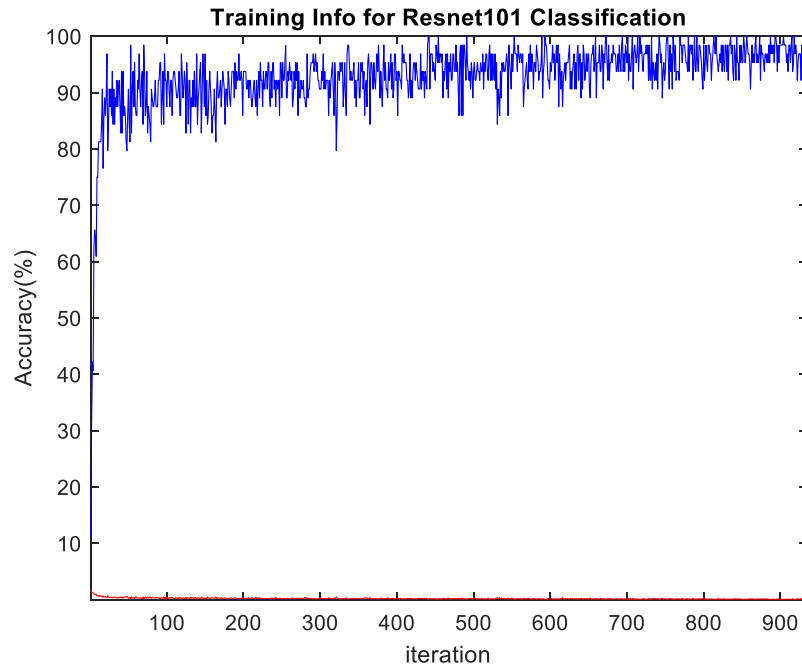


Figure 13: Accuracy and loss function obtained in the Resnet-101 training phase

In Figure 14, the confusion matrix obtained from the test data after training the Resnet-101 model is given.

True Class	1	15	4		2
	2	6	196	4	65
	3		1	1107	3
	4	5	24	5	1568
		1	2	3	4
		Predicted Class			

Figure 14: Confusion matrix for Resnet-101

The performance criteria obtained from Resnet-101 are given in Table 9. The model achieved an accuracy rate of 96.04% in total performance. Table 7 shows that high success rates were obtained for the 3rd and 4th grades. When the class size is taken into account, the success rates of the other classes are satisfactory and promising.

Class	Recall	Precision	F-Score
1	0.5769	0.7142	0.6382
2	0.8711	0.7232	0.7902
3	0.9919	0.9964	0.9941
4	0.9572	0.9787	0.9678
Weighted Avg.	0.9596	0.9603	0.9592

Table 9: Evaluation metrics for Resnet-101

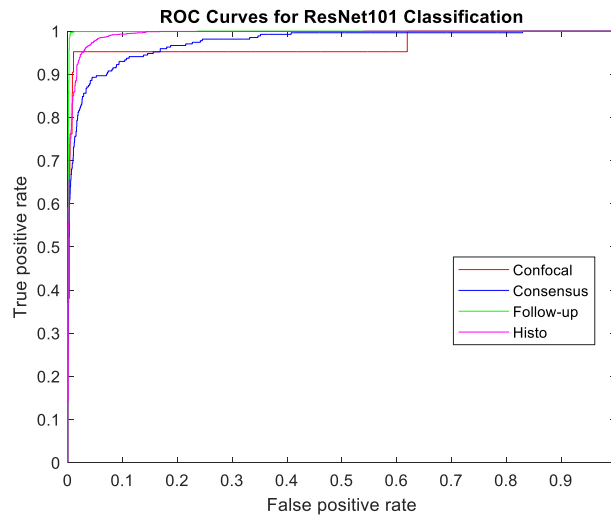


Figure 15: ROC curve for the Resnet-101 classification

The ROC curve for the Resnet-101 model is shown in Figure 15. When the ROC curve is analyzed, a successful classification is observed because all classification curves are close to the upper margin. The most unsuccessful classification among the classes was obtained in the “Consensus” class.

5.5 Comparative results

Because of the experimental studies, the best accuracy was obtained with the Resnet-101 model, as shown in Table 10.

Model	Accuracy (%)
VGG19	91.71
Resnet-50	92.51
Resnet-101	96.04
InseptionResnetV2	90.12

Table 10: Comparative results

The features of the ResNet-101 model, which achieved the highest accuracy, were evaluated using different classifiers: 3-layer Fully Connected Network (FCN), Support Vector Machine (SVM), K-Nearest Neighbour (KNN), and Gradient Boosting Machine (GBM) methods. The obtained results are presented in Table 11.

Model	Accuracy (%)
Resnet-101+3FCN	93.84
Resnet-101+SVM	88.68
Resnet-101+KNN	87.55
Resnet-101+GBM	86.45
Resnet-101	96.04

Table 11: Resnet-101 features with different classifiers

When Table 11 is examined, the highest accuracy is achieved with the single-layer FCN found in the original structure of the ResNet-101 architecture.

6 Conclusions

Skin cancer is a common and serious disease that can cause death if not diagnosed and treated in time. Today, artificial intelligence models have started to be used quite frequently in the field of medicine due to the increase in the effectiveness of artificial intelligence algorithms and the success of image recognition algorithms. When studies in the literature are examined, it is observed that researchers generally focus on classifying the types of skin cancer. However, skin cancer verification methods play an important role in creating a treatment planning process suitable for the patient after the cancer type is detected. Skin cancer confirmation methods provide important information to determine the stage, potential for spread, and other characteristics of the disease. For this purpose, skin cancer verification methods were classified using dermoscopic HAM10000 dataset images. Within the scope of the study, 4 different transfer learning models were used: Resnet-50, Resnet-101, VGG19, and models. The highest accuracy rate among the models was obtained with Resnet-101 (96.04%). Simultaneously, the models were subjected to maximum voting, and an accuracy rate of 93.14% was achieved. When the models were examined, it was seen that the success rates of this class were relatively low due to the lack of data for the 1st and 2nd grades. In the future, achieving more successful classification can be possible by increasing the data size of each class to a sufficient level and incorporating methods to address the problem of data imbalance. Thus, a system that performs well across all skin cancer datasets can be obtained.

References

- [Alam et al., 2022] Alam, T. M., Shaukat, K., Khan, W. A., Hameed, I. A., Almuqren, L. A., Raza, M. A., Aslam, M., & Luo, S. (2022). An Efficient Deep Learning-Based Skin Cancer Classifier for an Imbalanced Dataset. *Diagnostics*, 12(9), 2115.
- [Ali and Deserno, 2012] Ali, A. R. A., & Deserno, T. M. (2012). A systematic review of automated melanoma detection in dermoscopic images and its ground truth data. *Medical Imaging 2012: Image Perception, Observer Performance, and Technology Assessment*, 8318, 421-431.
- [Argenziano and Soyer, 2001] Argenziano, G., & Soyer, H. P. (2001). Dermoscopy of pigmented skin lesions—a valuable tool for early. *The lancet oncology*, 2(7), 443-449.

- [Bibi et al, 2023] Bibi, S., Khan, M. A., Shah, J. H., Damaševičius, R., Alasiry, A., Marzougui, M., Alhaisoni, M., & Masood, A. (2023). MSRNet: Multiclass skin lesion recognition using additional residual block-based fine-tuned deep models, information fusion, and best feature selection. *Diagnostics*, 13(19), 3063.
- [Bozkurt, 2023] Bozkurt, F. (2023). Skin lesion classification on dermatoscopic images using effective data augmentation and pre-trained deep learning approach. *Multimedia Tools and Applications*, 82(12), 18985-19003.
- [Brinker et al, 2018] Brinker, T. J., Hekler, A., Utikal, J. S., Grabe, N., Schadendorf, D., Klode, J., Berking, C., Steeb, T., Enk, A. H., & Von Kalle, C. (2018).. Skin cancer classification using convolutional neural networks: Systematic review. *Journal of Medical Internet Research*, 20(10), e11936.
- [Chen et al., 2017] Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4), 834-848.
- [Datta et al., 2021] Datta, S. K., Shaikh, M. A., Srihari, S. N., & Gao, M. (2021). Soft Attention Improves Skin Cancer Classification Performance. In *Interpretability of Machine Intelligence in Medical Image Computing, and Topological Data Analysis and Its Applications for Medical Data* (pp. 13-23). Springer, Cham.
- [Demir, 2021] Demir, F. (2021). Derin Öğrenme Tabanlı Yaklaşımla Kötü Huyllu Deri Kanserinin Dermatoskopik Görüntülerden Saptanması. *Fırat Üniversitesi Mühendislik Bilimleri Dergisi*, 33(2), 617-624.
- [Fabbrocini et al., 2010] Fabbrocini, G., Triassi, M., Mauriello, M. C., Torre, G., Annunziata, M. C., De Vita, V., Pastore, F., D'Arco, V., & Monfrecola, G. (2010). Epidemiology of skin cancer: role of some environmental factors. *Cancers*, 2(4), 1980-1989.
- [Fabbrocini et al., 2011] Fabbrocini, G., De Vita, V., Pastore, F., D'Arco, V., Mazzella, C., Annunziata, M. C., Cacciapuoti, S., Mauriello, M. C., & Monfrecola, A. (2011). Teledermatology: from prevention to diagnosis of nonmelanoma and melanoma skin cancer. *International journal of telemedicine and applications*, 2011.
- [Garg et al., 2021] Garg, R., Maheshwari, S., & Shukla, A. (2021). Decision support system for detection and classification of skin cancer using CNN. In *Innovations in Computational Intelligence and Computer Vision* (pp. 578-586). Springer, Singapore.
- [Haenssle et al., 2018] Haenssle, H. A., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., Kalloo, A., Ben Hadj Hassen, A., Thomas, L., Enk, A., & Uhlmann, L. (2018). Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of oncology*, 29(8), 1836-1842.
- [Hameed et al., 2021] Hameed, A., Umer, M., Hafeez, U., Mustafa, H., Sohaib, A., Siddique, M. A., & Madni, H. A. (2021). Skin lesion classification in dermoscopic images using stacked Convolutional Neural Network. *Journal of Ambient Intelligence and Humanized Computing*, 1-15.
- [He et al., 2016a] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

- [He et al., 2016b] He, K., Zhang, X., Ren, S., & Sun, J. (2016, October). Identity mappings in deep residual networks. In *European conference on computer vision* (pp. 630-645). Springer, Cham.
- [Hoang et al., 2022] Hoang, L., Lee, S. H., Lee, E. J., & Kwon, K. R. (2022). Multiclass Skin Lesion Classification Using a Novel Lightweight Deep Learning Framework for Smart Healthcare. *Applied Sciences*, 12(5), 2677.
- [Huang et al., 2021] Huang, H. W., Hsu, B. W. Y., Lee, C. H., & Tseng, V. S. (2021). Development of a light-weight deep learning model for cloud applications and remote diagnosis of skin cancers. *The Journal of Dermatology*, 48(3), 310-316.
- [Hussain et al., 2023] Hussain, M., Khan, M. A., Damaševičius, R., Alasiry, A., Marzougui, M., Alhaisoni, M., & Masood, A. (2023). SkinNet-INIO: multiclass skin lesion localization and classification using fusion-assisted deep neural networks and improved nature-inspired optimization algorithm. *Diagnostics*, 13(18), 2869.
- [Jaworek-Korjakowska et al., 2019] Jaworek-Korjakowska, J., Kleczek, P., & Gorgon, M. (2019). Melanoma thickness prediction based on convolutional neural network with VGG-19 model transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (pp. 0-0).
- [Khan et al., 2021a] Khan, M. A., Zhang, Y. D., Sharif, M., & Akram, T. (2021). Pixels to classes: intelligent learning framework for multiclass skin lesion localization and classification. *Computers & Electrical Engineering*, 90, 106956.
- [Khan et al., 2021b] Khan, M. A., Sharif, M., Akram, T., Damaševičius, R., & Maskeliūnas, R. (2021). Skin lesion segmentation and multiclass classification using deep learning features and improved moth flame optimization. *Diagnostics*, 11(5), 811.
- [Khan et al., 2021c] Khan, M. A., Muhammad, K., Sharif, M., Akram, T., & de Albuquerque, V. H. C. (2021). Multi-class skin lesion detection and classification via teledermatology. *IEEE Journal of Biomedical and Health Informatics*, 25(12), 4267-4275.
- [Kittler et al., 2002] Kittler, H., Pehamberger, H., Wolff, K., & Binder, M. J. T. I. O. (2002). Diagnostic accuracy of dermoscopy. *The lancet oncology*, 3(3), 159-165.
- [Krizhevsky et al., 2017] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6), 84-90.
- [Lan et al., 2022] Lan, Z., Cai, S., He, X., & Wen, X. (2022). FixCaps: An improved capsules network for diagnosis of skin cancer. *IEEE Access*, 10, 76261-76267.
- [Litjens et al., 2017] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical image analysis*, 42, 60-88.
- [Maqsood and Damaševičius, 2023] Maqsood, S., & Damaševičius, R. (2023). Multiclass skin lesion localization and classification using deep learning based features fusion and selection framework for smart healthcare. *Neural networks*, 160, 238-258.
- [Nami et al., 2012] Nami, N., Giannini, E., Burrioni, M., Fimiani, M., & Rubegni, P. (2012). Teledermatology: state-of-the-art and future perspectives. *Expert Review of Dermatology*, 7(1), 1-3.
- [Park et al., 2004] Park, S. H., Goo, J. M., & Jo, C. H. (2004). Receiver operating characteristic (ROC) curve: practical review for radiologists. *Korean journal of radiology*, 5(1), 11-18.

- [Popescu et al., 2022] Popescu, D., El-Khatib, M., & Ichim, L. (2022). Skin Lesion Classification Using Collective Intelligence of Multiple Neural Networks. *Sensors*, 22(12), 4399.
- [Rahi et al., 2021] Rahi, M. M. I., Khan, F. T., Khan, M. T., Ullah, A. A., & Alam, M. G. R. (2021, August). Transfer Learning Approach and Analysis for Skin Cancer Detection. In 2021 International Conference on Science & Contemporary Technologies (ICSCT) (pp. 1-6). IEEE.
- [Ren, 2023] Ren, G. (2023). Monkeypox Disease Detection with Pretrained Deep Learning Models. *Information Technology and Control*, 52(2), 288-296.
- [Simonyan et al., 2014] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
- [Srinivasu et al., 2021] Srinivasu, P. N., SivaSai, J. G., Ijaz, M. F., Bhoi, A. K., Kim, W., & Kang, J. J. (2021). Classification of skin disease using deep learning neural networks with MobileNet V2 and LSTM. *Sensors*, 21(8), 2852.
- [Stateczny et al., 2022] Stateczny, A., Uday Kiran, G., Bindu, G., Ravi Chythanya, K., & Ayyappa Swamy, K. (2022). Spiral Search Grasshopper Features Selection with VGG19-ResNet50 for Remote Sensing Object Detection. *Remote Sensing*, 14(21), 5398.
- [Szegedy et al., 2017] Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. A. (2017, February). Inception-v4, inception-resnet and the impact of residual connections on learning. In Thirty-first AAAI conference on artificial intelligence.
- [Thurnhofer-Hemsi and Domínguez, 2021] Thurnhofer-Hemsi, K., & Domínguez, E. (2021). A convolutional neural network framework for accurate skin cancer detection. *Neural Processing Letters*, 53(5), 3073-3093.
- [Tschandl et al., 2018] Tschandl, P., Rosendahl, C., & Kittler, H. (2018). The HAM10000 dataset, a large collection of multi-source dermoscopic images of common pigmented skin lesions. *Scientific data*, 5(1), 1-9.
- [Wang et al., 2021] Wang, J., He, X., Faming, S., Lu, G., Cong, H., & Jiang, Q. (2021). A Real-Time Bridge Crack Detection Method Based on an Improved Inception-Resnet-v2 Structure. *IEEE Access*, 9, 93209-93223.
- [Yu et al., 2021] Yu, H. Q., & Reiff-Marganiec, S. (2021). Targeted Ensemble Machine Classification Approach for Supporting IoT Enabled Skin Disease Detection. *IEEE Access*, 9, 50244-50252.